Statistical Methods & Applications (2024) 33:753–781
https://doi.org/10.1007/s10260-024-00757-x

ORIGINAL PAPER

Performance evaluation of nursing homes using finite mixtures of logistic models and M-quantile regression for binary data

G. De Novellis[1] · M. Doretti[2] · G. E. Montanari[3] · M. G. Ranalli[3] · N. Salvati[4]

Accepted: 11 June 2024 / Published online: 15 July 2024

Abstract

Evaluating the performance of health care institutions is of paramount interest and it is often conducted using generalized linear mixed models. In this paper, we focus on the evaluation of Nursing Homes for elderly residents in a region of Italy and concentrate on binary outcomes (death and worsening). We propose to use a routinely assessed covariate such as the Resource Utilization Group to account for case-mix. We fit finite mixtures of logistic models to check the assumption of normality of the random effects in the generalized linear mixed model approach and to obtain a clustering of the Nursing Homes with respect to their performance. Since the distribution of the random effects is very skew, we propose to use scores based on robust M-Quantile regression for binary data and estimate their standard error using block-bootstrap. A sensitivity analysis is also conducted to evaluate the assumption of missing at random for non-observed data on discharged residents.

Keywords Case-mix · Mixed logistic model · Bootstrap · Resource utilization groups

✉ M. G. Ranalli
maria.ranalli@unipg.it

[1] SDA Bocconi, School of Management, Milano, Italy

[2] Department of Statistics, Computer Science, and Applications, Università degli Studi di Firenze, Florence, Italy

[3] Department of Political Science, Università degli Studi di Perugia, Perugia, Italy

[4] Department of Economics and Management, Università di Pisa, Pisa, Italy

1 Introduction

As it is well-known, in Western countries Nursing Home (NH) care services are often supplied by private facilities, that typically get access to specific financial programs implemented by national or regional governments. Within this framework, equitable cost-based reimbursement schemes usually account for *case-mix*, that is, for the overall level of clinical complexity that each NH has to cope with (Brizioli et al. 2003). Indeed, NH residents typically find themselves in quite different health conditions when entering the facilities.

A well known approach to measure case-mix consists in classifying NH residents into Resource Utilization Groups (RUGs; Schneider et al. 1988; Fries et al. 1994). The underlying idea is that residents belonging to the same RUG require approximately the same amount of resources to take care of them. Such a classification method was found to be quite effective in explaining the variability in NH managing costs (Ikegami et al. 1994; Fries et al. 1994), and it is still used in various public health care contexts; see for example Punelli and Williams (2013) and Broussard and Reiter (2020). The RUG classification version currently in use is RUG III (Fries et al. 1994).

The importance of adjusting for case-mix extends to another area of primary interest: the evaluation of care facilities' performance (Berlowitz et al. 1996). In detail, when case-mix related factors are also associated with the outcome variables (like in most cases), some degree of confounding arises, and adjusting for such factors turns out to be necessary in order to make fair comparisons (Wray et al. 1997).

In this paper, we build upon this framework and exploit the RUG classification for adjustment purposes, when comparing the performances of a set of NHs with respect to relevant outcomes. Specifically, we fit a set of statistical models to data collected on residents hosted by a group of NHs in Umbria, a region of central Italy. Our overall aim is assessing NH performance in relative terms, that is, taking their average level as benchmark. For all these models, case-mix adjustment is addressed by including residents' RUG as a covariate.

Although embedded in a cross-sectional framework, our analyses take the same perspective as those implemented for NH performance evaluation with longitudinal data (Bartolucci et al. 2009; Montanari et al. 2018; Montanari and Doretti 2019). Indeed, NH performance remains framed in terms of ability to preserve as much as possible residents' health status. In this respect, we consider two binary outcomes of interest: resident death and resident worsening, both one year after baseline. As detailed in Sect. 2, worsening is defined in terms of RUGs as well.

Like in the more developed literature on hospital evaluation (see, e.g., Grieco et al. 2012; Berta et al. 2016; Berta and Vinciotti 2019), refined performance assessment approaches build upon mixed effect models (Goldstein 2011). In particular, in our setting this corresponds to fitting Logistic Mixed Models (LMMs) with a Gaussian random effect at the NH level. Such an effect can be used as a performance marker, provided that the covariates included in the fixed part of the model properly account for the case-mix. For the data at hand, we start by fitting

LMMs but then, motivated by evidence against the normality assumption for the random effect distribution, we explore two more robust alternatives: Finite Mixtures of Logistic Models (FMLMs) and M-quantile regression for binary data.

In FMLMs, a discrete distribution for the NH-specific random effects is estimated from the observed data via a Non-Parametric Maximum Likelihood approach (NPML; Simar 1976; Laird 1978; Lindsay 1983a, b). This approach offers a number of advantages. First, it allows to avoid unverifiable assumptions on the random effect distribution, so that asymmetries or other departures from normality can be easily accommodated. In this sense, FMLMs are more robust to misspecification of the distribution of the random effects than LMMs. Also, FMLMs provide a data-driven method to cluster NHs according to their probability of belonging to the different mixture components, identified by the support points for the discrete random effect. The resulting clusters have a clear interpretation in terms of NH performance.

With regard to M-quantile regression, we rely on the approach of Kokic et al. (1997), developed for comparing business production. In general, M-quantile regression (Breckling and Chambers 1988) provides a quantile-like generalization of robust M-regression. It can also be seen as the robust version of expectile regression proposed by Newey and Powell (1987). While in linear regression we model the expected value of the conditional distribution of the outcome given the covariates, with quantile regression (Koenker and Bassett 1978) we model the quantiles of this conditional distribution. On the other hand, with M-quantile regression we model a robustification of the expectiles of this conditional distribution using M-estimation via Huber-type influence functions. An extension of this approach to binary responses is introduced in Chambers et al. (2016). In this model, every observation lies on one of the estimated M-quantile regression hyperplanes: this provides a score between 0 and 1 that corresponds to the M-quantile of the distribution of the response variable each observation is estimated to belong to, conditional on the covariates included in the model. Suitably averaging these scores for units belonging to the same NH provides an alternative measure of performance. Importantly, such a measure does not depend on the level of the covariates, that is, on case-mix. Bootstrap is used to obtain a measure of uncertainty of the NHs' average scores. There are many proposals in the literature that deal with quantile regression for binary data. See e.g. Benoit and Van den Poel (2012), Kordas (2006), Aristodemou et al. (2019). However, the focus here is on exploring the use of the M-quantile coefficient for performance evaluation, as M-quantiles are a direct extension of the logistic regression model and look at (possibly robust) expectiles rather than quantiles.

The paper is organized as follows. Section 2 describes the dataset at hand, while Sect. 3 illustrates LMMs and the two alternative methodologies considered to obtain a performance measure for NHs in this context: FMLMs and M-quantile regression for binary data. Section 4 details the results of the application of these methods to the data and Sect. 5 provides conclusions and directions for future research.

2 Data

The healthcare division of the regional government of Umbria (Italy) routinely collects data on the residents in the NHs of the region for monitoring purposes. The analysis of this NH system is based upon data coming from the Long Term Care Facilities (LTCF) questionnaires (Hirdes et al. 2008; Kim et al. 2015). The administration of LTCF questionnaires is implemented within an internationally validated protocol named *Suite interRAI* (Carpenter and Hirdes 2013). Such a protocol is adopted by many other regional governments in Italy. The questionnaires are filled by the NH staff and investigate several aspects of NH residents' health status as well as the medical treatments undertaken.

Using information from LTCF questionnaires, residents are classified according to the RUG system introduced in Sect. 1. Specifically, the underlying algorithm assigns residents to 44 distinct groups. These are partitioned into six macro-groups: Rehabilitation (not present in Umbrian NHs by law), Extensive services, Special care, Clinically complex, Impaired cognition, Behavior problems, Reduced physical functions. Attached to each group comes a weight, which is a numerical proxy of the conventional amount of resources required to take care of a resident in that RUG. Such a weight is a pure number ranging in our data from 0.52 to 1.86.

The LTCF questionnaires are administered to NH residents every six months and whenever a significant change in the health conditions is observed. Thus, a sort of longitudinal dataset is available. This allows to single out the set of residents hosted by the NHs at a chosen baseline data, as well as to observe their condition after a suitable time interval. A comparison between the final and the baseline observation can be performed to define some statistical indicators measuring NH ability to preserve their residents' health status and/or to avoid their worsening over time.

In this paper, we consider residents hosted by the Umbrian NHs on January 1st, 2018. For each of them, we observe the RUG associated to the last LTCF questionnaire filled before that date. Henceforth, this group will be referred to as the Initial RUG (IRG), while the corresponding macro-group it belongs to will be denoted by IMRG. For the same residents, we also observe the RUG associated to the last questionnaire available before January 1st, 2019, referred to as Final RUG (FRG). Since it can happen that in the meantime a resident dies or gets discharged, we add two additional groups to the FRGs: the "Death" and the "Discharge" groups. This approach reflects the fact that death and discharge are proper outcomes of the process of interest, rather than simple causes of missing values in the data.

The available dataset includes $n = 1551$ residents. For each resident, information on age and gender is available together with NH membership, IRG, FRG and the corresponding RUG weights. Furthermore, we compute two binary outcomes: (*i*) a death indicator, taking the value 1 when the FRG is the "Death" group, and (*ii*) a worsening indicator, taking the value 1 when the difference between the FRG weight

Table 1 Main descriptive statistics of NH characteristics at baseline

Variable	Minimum	Q1	Q2	Mean	Q3	Maximum
Nr. of residents	3	19	24	33	46	80
Mean Age	74.0	81.4	84.0	83.7	85.6	94.7
% of Males	0.0	22.5	28.6	29.0	36.0	59.1
% of Deaths*	5.3	15.0	20.8	21.6	26.3	56.3
% of Worsened*	5.3	21.4	30.0	30.2	35.4	75.0
% of Discharged*	0.0	0.0	0.0	2.6	4.3	40.0
Mean IRG Weight	0.70	0.79	0.86	0.86	0.93	1.05
% of Special care	0.0	3.5	5.9	7.1	10.6	21.4
% of Clinically complex	0.0	8.3	15.4	16.5	22.9	41.3
% of Impaired cognition	0.0	1.2	5.6	7.7	11.4	28.6
% of Reduced physical function	14.3	29.7	36.4	38.6	48.3	71.4
% of Behavior problems	0.0	0.0	0.0	1.6	2.7	10.0
% of Extensive services	0.0	0.0	2.1	4.3	5.9	26.3

*One year after baseline

and the IRG weight is positive, denoting an increase in the care burden required by the resident, or when the FRG is the "Death" group. With regard to discharge, its determinants are various and not known from the data, so a well-established and univocal relationship with resident health status and care burden cannot be postulated.

The NH residents are hosted in $m = 47$ NHs. Table 1 reports summary statistics at the NH level. Each NH hosts on average 33 residents, though with some degree of variability. The mean of the NH average age of residents is 83.5 years, with standard deviation 10.3. The proportion of females is slightly greater than 70% on average. The average death rate is around 22%, ranging from 5.3% to 56.3% across NHs. Similarly, the average discharge rate is around 3%, ranging from 0% to 40% across NHs. Note that 29 out of 47 NHs do not present resident discharge.

The rather relevant variability in death rates across NHs can be due to case-mix, on the one hand, and to the quality of care provided by the NH, on the other hand. Therefore, through the statistical models mentioned in Sect. 1, we propose to estimate the NH effect on the probability of death or worsening of the residents, under the assumption that the latter reflects NH ability to preserve residents' health, after properly accounting for the case-mix.

3 Methods

Let y_{ij} be a binary response variable for unit $j = 1, \dots, n_i$ belonging to NH $i = 1, \dots, m$, and let $\boldsymbol{x}_{ij} = (x_{ij1}, \dots, x_{ijp})^T$ denote the corresponding vector of explanatory variables. Consider the case in which this binary response is modeled via a logistic mixed model with random intercepts u_i in order to obtain a measure of the NH effect on the response variable. Then, conditional on the NH-specific random effect u_i, responses for units in the i-th NH are assumed to be independent Bernoulli random variables with success probability $p_{ij} = E(y_{ij} \mid u_i; \boldsymbol{x}_{ij})$ described by the following model

$$\log \frac{p_{ij}}{1 - p_{ij}} = \beta_0 + \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + u_i. \tag{1}$$

Responses are assumed to be independent conditional on the random variable u_i; this is usually referred to as the local independence assumption. The corresponding likelihood function is

$$L(\boldsymbol{\Phi}) = \prod_{i=1}^{m} \left\{ \int_{\mathbb{R}} \prod_{j=1}^{n_i} f(y_{ij} | u_i; \boldsymbol{x}_{ij}) f(u_i) \, \mathrm{d}\, u_i \right\}, \tag{2}$$

where $f(y_{ij}|u_i;\boldsymbol{x}_{ij})$ is the Bernoulli distribution for the outcome, $f(u_i)$ is the random coefficient distribution, and $\boldsymbol{\Phi}$ denotes the global set of parameters. The terms u_i, $i = 1, \dots, m$, are meant to model unobserved NH-specific heterogeneity common to each resident (lower-level unit) within the same i-th NH (upper-level unit) and, in particular, to account for dependence between responses recorded within the same NH. Usually $f(\cdot)$ is taken from a parametric distribution, with the Gaussian being the most popular choice: $u_i \sim N(0, \sigma_u^2)$, $i = 1, \dots, m$. In the general case, the integral defining the likelihood cannot be analytically computed. For maximum likelihood estimation, the integral can be approximated using (adaptive) Gaussian Quadrature or Laplace approximation (see e.g. Pinheiro and Bates 1995), Monte Carlo EM methods (see e.g. McCulloch 1997), Penalized Quasi-Likelihood (PQL) approaches or Taylor linearization methods (Breslow and Clayton 1993). To overcome the issue, Jiang (1998) suggested to derive estimates by exploiting the method of moments.

For performance evaluation of NHs, prediction of the vector $\boldsymbol{u} = (u_1, \dots, u_m)^T$ of random effects is essential. For known $\boldsymbol{\Phi}$, the Best Predictor of $\boldsymbol{u}$ in terms of minimum MSE is its conditional expectation given by

$$\tilde{\boldsymbol{u}}(\boldsymbol{\Phi}) = E_{u|y}(\boldsymbol{u}|\boldsymbol{y}) = \int_{\mathbb{R}^m} \boldsymbol{u} f(\boldsymbol{u}|\boldsymbol{y}) \, \mathrm{d}\, \boldsymbol{u},$$

where

$$f(\boldsymbol{u}|\boldsymbol{y}) = \frac{\prod_{i=1}^{m} \prod_{j=1}^{n_i} f(y_{ij}|u_i;\boldsymbol{x}_{ij}) f(u_i)}{\prod_{i=1}^{m} \int_{\mathbb{R}} \prod_{j=1}^{n_i} f(y_{ij}|u_i;\boldsymbol{x}_{ij}) f(u_i) \, \mathrm{d}\, u_i}.$$

This suggests the predictor $\hat{\boldsymbol{u}} = \tilde{\boldsymbol{u}}(\hat{\boldsymbol{\Phi}})$, where $\hat{\boldsymbol{\Phi}} = (\hat{\beta}_0, \hat{\boldsymbol{\beta}}^T, \hat{\sigma}_u)^T$ are suitable estimates. The computation of the Best Predictor requires the evaluation of integrals. These can be approximated using again (adaptive) Gaussian Quadrature or the Laplace approximation. Penalized iteratively reweighted least squares (PIRLS) and PQL methods directly estimate $\boldsymbol{u}$ (see e.g. Saei and Chambers 2003), but they may provide inconsistent model parameter estimates. Monte Carlo EM methods produce an estimate of β_0 and of $\boldsymbol{\beta}$, and since one generates a sample of the $\boldsymbol{u}$'s from the distribution of $\boldsymbol{u}$ given the data, the mean of these samples provides an estimate of $\boldsymbol{u}$.

Similarly, in a Bayesian setting, MCMC provides a sample of the $\boldsymbol{u}$'s, and the mean of this sample yields an estimate of $\boldsymbol{u}$.

As long as the covariates account properly for the case-mix, the random effects u_i can be considered as measures of the performance of the NHs (Goldstein 2011). This approach has been used extensively for performance evaluation in general and of health care facilities, in particular (Berta et al. 2016; Grieco et al. 2012). In the following subsections, we illustrate two alternative approaches to obtain a measure of performance for binary outcomes in the abovementioned framework.

3.1 Finite mixtures of logistic regression models

As a first alternative approach we use FMLMs to relax the assumption of normality for the distribution of the random effects. Indeed, the approach is more general and avoids any parametric assumption on the distribution $f(u_i)$. Rather than specifying a parametric distribution for the random effects, we may leave it unspecified and approximate it by using a discrete distribution on $G < m$ locations $\{u_1, \dots, u_G\}$, with associated probabilities defined by $\pi_k = \Pr(u_i = u_k)$, $i = 1, \dots, m$ and $k = 1, \dots, G$. That is, $u_i \sim \sum_{k=1}^{G} \pi_k \delta_{u_k}$ where δ_θ is a one-point distribution putting a unit mass at θ. In this case, the likelihood in Eq. (2) reduces to

$$L(\boldsymbol{\Phi}) = \prod_{i=1}^{m} \left\{ \sum_{k=1}^{G} \prod_{j} f(y_{ij} | u_k; \boldsymbol{x}_{ij}) \pi_k \right\} =: \prod_{i=1}^{m} \left\{ \sum_{k=1}^{G} \prod_{j} f_{ijk} \pi_k \right\}. \tag{3}$$

Equation (3) resembles the likelihood function for a finite mixture of Bernoulli distributions, where $\boldsymbol{\Phi} = (\boldsymbol{\beta}, u_1, \dots, u_G, \pi_1, \dots, \pi_G)$, and f_{ijk} is the distribution of the response variable for the j-th unit in the i-th NH when the k-th component of the finite mixture, $k = 1, \dots, G$, is considered.

The above model may be embedded into the class of semi-parametric mixed-effect models, that was recently extended to deal with continuous (Masci et al. 2019), categorical (Masci et al. 2022) and time-to-event outcomes (Gasperoni et al. 2020). With this regard, it can be seen as a semi-parametric, discrete approximation to a fully parametric, possibly continuous, distribution for the random coefficients. The seminal papers by Aitkin (1996, 1999) establish a connection between mixed effect models and finite mixtures, by exploiting the theory of Non Parametric Maximum Likelihood estimation of a mixing distribution, see Laird (1978). It can be also thought of as a model-based clustering approach, where the population of interest is assumed to be divided into G homogeneous sub-populations which differ for the values of the intercept, as in Wedel et al. (1993). This feature of the approach turns out to be particularly useful with the application at hand, as it provides a data-driven clustering of the NHs that accounts for the level of the covariates and is, therefore, associated with their performance. In addition, it provides a useful tool to assess whether the assumption of Normality for the random effects is plausible in LMMs.

Compared to a fully parametric framework, the finite mixture approach is less parsimonious, since the number of unknown parameters to be estimated is higher than in the corresponding parametric model. In fact, locations u_k and masses π_k, $k = 1, \dots, G$, are unknown parameters, as it is G, which is usually treated as fixed and chosen through appropriate penalized likelihood criteria. While we also follow this strategy, it is worth to mention the recent proposal of Ragni et al. (2023), based on an adaptive procedure that, at each iteration of the estimation algorithm, merges two mixtures components if the difference between their location parameters is not statistically significant.

When using discrete random coefficients, the regression model in Eq. (1) for the k-th component of the mixture can be expressed as follows:

$$\theta_{ijk} =: \log \frac{p_{ijk}}{1 - p_{ijk}} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + u_k, \tag{4}$$

where u_k is now a proper random intercept.

Given the model assumptions, the score function can be written as the posterior expectation of the score function corresponding to a standard logistic model:

$$\begin{aligned} S(\boldsymbol{\Phi}) &= \frac{\partial \log[L(\boldsymbol{\Phi})]}{\partial \boldsymbol{\Phi}} = \frac{\partial \ell(\boldsymbol{\Phi})}{\partial \boldsymbol{\Phi}} \\ &= \sum_{i=1}^{m} \sum_{k=1}^{G} \left(\frac{f_{ik}\pi_k}{\sum_l f_{il}\pi_l} \right) \sum_j \frac{\partial \log f_{ijk}}{\partial \boldsymbol{\Phi}} =: \sum_{i=1}^{m} \sum_{k=1}^{G} \tau_{ik} \sum_j \frac{\partial \log f_{ijk}}{\partial \boldsymbol{\Phi}}, \end{aligned} \tag{5}$$

where f_{ik} is the joint conditional probability for the observed responses in the i-th NH and the k-th component, i.e.

$$f_{ik} = \exp \left\{ \sum_{j=1}^{n_i} \left[y_{ij}\theta_{ijk} - \log(1 + e^{\theta_{ijk}}) \right] \right\},$$

and the weights

$$\tau_{ik} = \frac{f_{ik}\pi_k}{\sum_l f_{il}\pi_l} \tag{6}$$

represent the posterior probabilities of component membership. Equating (5) to zero gives likelihood equations that are essentially weighted sums of the likelihood equations for a standard logistic model, with weights τ_{ik}. Maximum likelihood estimation is based on the use of EM, or EM-type, algorithms (Dempster et al. 1977). The basic EM algorithm is defined by solving equations for a given set of the weights, and updating the weights as a function of the current parameter estimates.

Under this approach, the NH-specific effects can be obtained computing the posterior mean of the estimated location points as

$$\hat{u}_i = \sum_{k=1}^{G} (\hat{u}_k - \hat{\bar{u}}) \hat{\tau}_{ik}, \tag{7}$$

where $\hat{\bar{u}} = \sum_k \hat{u}_k \hat{\pi}_k$ is the overall intercept estimate.

3.2 M-quantile regression for binary data

The second alternative approach to LMM moves away from random effect modeling to obtain a performance score for NHs. In particular, it looks at the conditional distribution of the outcome given the covariates by means of M-Quantile regression models for binary data. In fact, LMM and FMLM focus on the expected value of the binary (Bernoulli) response and provide a measure of the performance of the NHs based on the predicted value of the random effects. In this section, we propose a measure of performance by looking at the different levels of the conditional distribution of y given the set of explanatory variables, by means of M-quantile regression for binary data (Chambers et al. 2016).

For a continuous response, quantile regression (Koenker and Bassett 1978) leads to a family of hyperplanes indexed by a real number $q \in (0, 1)$ representing the quantile of interest. For example, for $q = 0.05$ the quantile regression hyperplane separates the lowest 5% of the conditional distribution from the remaining 95%. In this sense, quantile regression can be considered as a generalization of median regression (Koenker and Bassett 1978), as expectile regression (Newey and Powell 1987) is a quantile-like generalization of standard mean regression. M-quantile regression (Breckling and Chambers 1988) integrates these concepts within a framework defined by a quantile-like generalization of regression based on influence functions (M-regression) or, similarly, by a robustification of expectile regression via influence functions.

The M-quantile of order q for the conditional density of a continuous outcome y is defined as the solution MQ_q that satisfies

$$\int_{\mathbb{R}} \psi_q\left(\frac{y - MQ_q}{\sigma_q}\right) f(y)\,\mathrm{d}\,y = 0, \tag{8}$$

where $\psi_q(t) = 2\psi(t)\{qI(t > 0) + (1 - q)I(t \leq 0)\}$, ψ is an influence function and σ_q is a measure of scale for $y - MQ_q$. When $\psi(t) = t$, MQ_q is the expectile of order q, while when $\psi(t) = \text{ sign }(t)$, MQ_q is the standard quantile of order q. A linear M-quantile regression model is the one for which the M-quantile of order q of the conditional distribution of y given $\boldsymbol{x}$ is such that

$$MQ_q(y \mid \boldsymbol{x};\psi) = \beta_{0q} + \boldsymbol{x}^T \boldsymbol{\beta}_q, \tag{9}$$

Estimates of β_{0q} and of $\boldsymbol{\beta}_q$ on the available data are obtained by solving the following set of estimating equations

$$\sum_{i=1}^{m} \sum_{j=1}^{n_i} \psi_q\left(\frac{y_{ij} - MQ_q(y_{ij}|\boldsymbol{x}_{ij};\psi)}{\sigma_q}\right)(1, \boldsymbol{x}_{ij}^T)^T = \mathbf{0}_{p+1}, \tag{10}$$

where the scale parameter is often estimated as $\hat{\sigma}_q = $ median $|y_{ij} - MQ_q(y_{ij}|\boldsymbol{x}_{ij};\psi)|/0.6745$. The influence function ψ is usually chosen to be the Huber loss function, $\psi(t) = tI(-c \leq t \leq c) + c \cdot \text{ sgn }(t)I(|t| > c)$, where c is a tuning constant. Provided that $c > 0$, estimates of $\boldsymbol{\beta}_q$ are obtained using

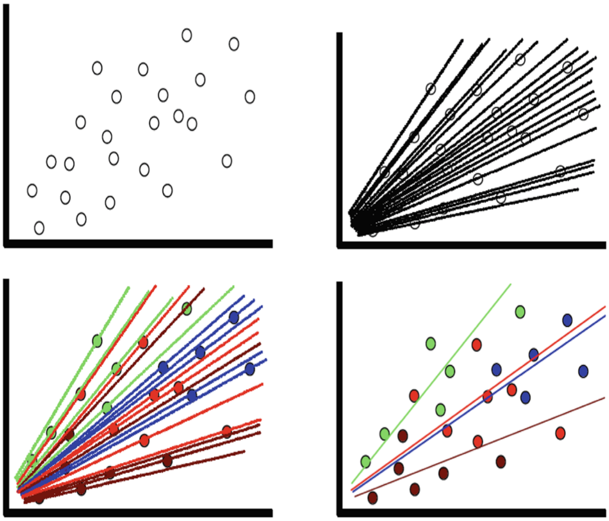


Fig. 1 An illustration of the rationale behind the performance score based on individual M-quantiles coefficients for a continuous response

Iteratively Weighted Least Squares (IWLS). In this case, the IWLS algorithm is known to guarantee convergence to a unique solution, see Kokic et al. (1997). The asymptotic theory for M-quantile regression with i.i.d. errors and fixed regressors can be derived from the results in Huber (1973) and is discussed in Breckling and Chambers (1988).

For a continuous outcome, Kokic et al. (1997) propose to develop a measure of business production performance within this framework. To illustrate the idea behind this approach, consider the case of simple M-quantile regression depicted in Fig. 1. Every observation in the scatterplot (first panel) can be thought of lying on a specific M-quantile regression line identified by its order q (second panel). In particular, the unit specific order q_{ij} is such that $y_{ij} = \beta_{0q_{ij}} + x_{ij1}\beta_{1q_{ij}}$. This is often called "M-quantile coefficient". An estimate of q_{ij} can be obtained by first fitting an ensemble of M-quantile regression lines for a fine grid of values $q \in (0;1)$ and, then, proceeding by interpolation of the two closest values.

If the observations were clustered (in the third panel observations are color coded according to the cluster they belong to) and if a cluster effect is indeed present in the data, then units belonging to the same cluster, i.e. to the same NH in this paper, should have a similar M-quantile coefficient as they should lie on a similar portion of the conditional distribution of the response given the covariates. In fact, if a hierarchical structure does explain part of the variability in the data, then we expect units within clusters defined by this hierarchy to have similar M-quantile coefficients, as for the green and the brown units in the illustration. Then, (fourth panel) a performance score can be obtained by suitably averaging the estimated M-quantile coefficients within the cluster as

$$\hat{q}_i = \sum_{j=1}^{n_i} \hat{q}_{ij}/n_i. \tag{11}$$

A similar approach has been used in Fiaschi et al. (2020) to develop an index of corporate wrongdoing, understood as firms' involvement in the number of controversies over universal human rights. Note that M-quantile regression with random effects is proposed in Tzavidis et al. (2015) to deal with longitudinal data on a continuous response. Therefore, it would be possible, by suitably extending Tzavidis et al. (2015) to the case of a binary response, to include the clustering structure directly into the regression model using random intercepts. Note that, however, using $\hat{q}_i$ as a performance score instead of an estimate of these random effects, a structure with random intercepts and random slopes can be automatically accounted for (see, e.g., Fig. 1), without the need to make assumptions on their distribution.

Now, since the quantile of order q of a binary variable is not unique, there is no clear definition of a quantile function in this case. Nonetheless, M-quantiles of a binary variable exist and are unique as long as the influence function ψ is continuous and monotone non-decreasing (Chambers et al. 2016). In particular, for a binary variable y with probability of success equal to p, Eq. (8) becomes

$$pq\psi\left(\frac{1-MQ_q}{\sigma_q}\right) - (1-p)(1-q)\psi\left(\frac{MQ_q}{\sigma_q}\right) = 0. \tag{12}$$

Note that, when $\psi(t) = t$ and $q = 0.5$, the solution to this estimating equation is $MQ_{0.5} = p$.

In the presence of explanatory variables, $MQ_q(y \mid \boldsymbol{x};\psi)$ can be specified as follows

$$MQ_q(y \mid \boldsymbol{x};\psi) = \frac{\exp(\beta_{0q} + \boldsymbol{x}^T\boldsymbol{\beta}_q)}{1+\exp(\beta_{0q} + \boldsymbol{x}^T\boldsymbol{\beta}_q)}. \tag{13}$$

Estimates of β_{0q} and of $\boldsymbol{\beta}_q$ can be obtained using the extension to the M-quantile case of the approach proposed by Cantoni and Ronchetti (2001) for M-estimation of parameters of a generalized linear model. This approach has also been used to develop M-quantile regression for counts in Tzavidis et al. (2015) and in Dreassi et al. (2014). The details for binary outcomes can be found in Chambers et al. (2016). Here, it suffices to notice that the approach is based on robustification of the maximum likelihood estimating equations for a logistic model where the influence function is applied to Pearson residuals. In this sense, no explicit distributional assumptions are made as quasi-likelihood is employed together with a link function so that only a working mean and a variance function for y are required. Analytic formulas for the estimation of the variance covariance matrix of $(\hat{\beta}_{0q}, \hat{\boldsymbol{\beta}}_q^T)^T$ are also provided in Chambers et al. (2016).

The definition of the M-quantile coefficient for binary data is more challenging than for a continuous response. Chambers et al. (2016) propose to define it as the the value q_{ij} for which

$$MQ_{q_{ij}}(y_{ij} \mid \boldsymbol{x}_{ij};\psi) = \frac{MQ_{0.5}(y_{ij} \mid \boldsymbol{x}_{ij};\psi) + y_{ij}}{2}.$$

When a logistic specification for $MQ_q(y \mid \boldsymbol{x};\psi)$ is used, such as in (13), this definition is equivalent to defining q_{ij} as the solution to

$$y^*_{ij} = \log\left[\frac{MQ_{q_{ij}}(y_{ij} \mid \boldsymbol{x}_{ij};\psi)}{1 - MQ_{q_{ij}}(y_{ij} \mid \boldsymbol{x}_{ij};\psi)}\right], \tag{14}$$

where $y^*_{ij} = \beta_{0q_{ij}} + \boldsymbol{x}^T_{ij}\boldsymbol{\beta}_{q_{ij}}$. Since relevant analytic expressions are intractable, bootstrap methods can be used to evaluate standard error and confidence intervals for scores obtained as averages of M-quantile coefficients q_{ij}.

4 Application to the evaluation of Umbrian NHs

We report here the results obtained with the proposed approaches applied to the dataset described in Sect. 2. Among the $n = 1,551$ residents hosted by the Umbrian NHs as of January 1st, 2018, there are 42 cases of discharge before January 1st, 2019. For these residents, the values of the outcome variables Death and Worsening are missing and the reason of discharge is not recorded. In the first part of our analysis, we have considered data from these 42 discharged residents as missing at random (MAR; Little and Rubin 2002) and the statistical models have been fitted on the $n = 1,509$ remaining residents. In Sect. 4.4 we run a sensitivity analysis to evaluate the appropriateness of this assumption. The variables Age and Initial RUG Weight (IRW) are inserted as continuous variables, while Gender and the Initial Macro RUG Group (IMRG) as categorical variables.

4.1 Results from logistic mixed models

We fit several LMMs to the binary response variables Death and Worsening using the `glmer` function of the `lme4` package in R (Bates et al. 2015). The NH categorical variable is inserted as fixed effect or random effect in order to evaluate which approach would best fit these data.

Model selection is guided by the Bayes Information Criterion (BIC) and the Akaike Information Criterion (AIC). To reduce computational burden, the fast PIRLS algorithm (`nAGQ=0` option of `glmer`) is used. We did not find sensible differences in parameter estimates obtained using the Laplace approximation (`nAGQ=1`) and with an increasing number of quadrature points (`nAGQ` between 2 and 5). Among the explicative variables, besides those already mentioned, we also consider the interaction between gender and age (Gender×Age), and nonparametric functions of age and of IRW—spline(Age) and spline(IRW)—as approximated via B-splines. Table 2 reports the values of BIC and AIC for a set of alternative models

Table 2 BIC and AIC for logistic (mixed) models for outcomes Death and Worsening

	Death		*Worsening*	
Models*	BIC	AIC	BIC	AIC
~ Gender + Age + IMRG + IRW + NHs	1848.06	1555.51	2166.27	1873.71
~ Gender×Age + IMRG + IRW, random=NHs	1585.75	1527.24	1919.27	1860.76
~ Gender + Age + IMRG + IRW, random=NHs	1580.88	1527.69	1914.50	1861.31
~ Gender + spline(Age) + IMRG + IRW, random=NHs	1588.19	1529.68	1921.80	1863.28
~ Gender + Age + IMRG + spline(IRW), random=NHs	1588.19	1529.68	1921.80	1863.28
~ Gender + Age + IMRG, random=NHs	1583.33	1535.45	1907.78	1859.90
~ Gender + Age + IRW, random=NHs	1547.93	1521.33	1879.00	**1852.40**
~ IRW, random=NHs	1569.16	1553.21	1896.15	1880.19
~ Gender + Age, random=NHs	1586.44	1565.16	1878.02	1856.74
~ Age + IRW, random=NHs	1542.90	1521.62	**1873.84**	1852.57
~ Gender + Age + IRW	**1541.26**	**1519.99**	1881.76	1860.49

The smallest value for each column is in bold

*random=NHs denotes random intercept models where the hierarchy is induced by NHs. Models in which this term is not present are standard logistic regression models

Table 3 LMM: parameter estimates for the outcomes Death and Worsening

	Death		Worsening	
Intercept	−6.319	***	−4.211	***
	(0.707)		(0.600)	
Gender (male)	0.235		0.199	
	(0.155)		(0.136)	
Age	0.042	***	0.034	***
	(0.007)		(0.006)	
IRW	1.627	***	0.554	**
	(0.239)		(0.221)	
Std. Dev of random effects $\hat{\sigma}_u$	0.185		0.344	***
χ^2 (Likelihood ratio test for significance of $\hat{\sigma}_u$)	0.652		10.085	**

Standard errors of estimates are in parentheses (significance codes: '.' 0.1, '*' 0.05, '**' 0.01, '***' 0.001)

for the outcome variable Death (first two columns) and Worsening (second two columns). The smallest value for each column is in bold.

Looking at BIC and AIC for the response Death, the smallest value suggests to select the model with Gender, Age, IRW, and without a random NH effect. For the response Worsening, the minimum value of BIC suggests to select the model with Age and IRW as fixed effects, and with the NH random effect, while the minimum value of AIC suggests to also add Gender among the fixed effects. We decide to select as our final working model for both outcome variables the one that includes additive fixed effects of Gender, Age and IRW and additive random effects for the

NH. The reason for this choice is that Gender is in general a characteristic of the residents felt to be important to account for by practitioners, while the NH random effect is kept also for the outcome Death for sake of comparison with results from the other approaches, FMLM and M-quantile regression.

The parameter estimates for the final selected models are shown in Table 3. We note a positive relation between the probability of Death and the values of Age and of IRW. Similar effects are observed for the probability of Worsening, even though they are less strong. The estimated standard deviation of the random effects is larger for Worsening, by this highlighting a stronger NH effect for this outcome. Indeed, the estimate of σ_u is significantly different from zero only for this outcome. In fact, the likelihood ratio test for a zero variance component is not significant for the outcome Death, even comparing it with the $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$ distribution as suggested in Pinheiro and Bates (2000, Sect. 2.4.1). This provides evidence of a first finding of the analysis, i.e. that there is no suggestion of a significant NH effect on Death. This is also supported by the information criteria that take smaller (larger) values for the model with only fixed effects for the outcome Death (Worsening). See last line of Table 2.

Since the distribution of the random effects is central to the analysis at hand, we further explore it by using FMLMs: using these models allows us to evaluate whether the assumption of Normality is suitable for it and whether the lack of a NH effect for the outcome Death is due to a miss-specification of such a distribution. The findings of this analysis are reported in the following section.

4.2 Results from finite mixtures of logistic regression models

We fit several FMLMs using non-parametric maximum likelihood (Aitkin 1996) as implemented in the `allvc` function of the `npmlreg` package in R (Einbeck et al. 2018). The covariates in the model are those selected above for the final chosen LMM, while the number of mixture components ranges from 1 to 10. To select such a number, we pair likelihood-based indices like AIC and BIC with another entropy-based measure, which accounts for the sharpness of the underlying NH clustering process. Indeed, NHs can be assigned, via their posterior probability distribution, to one of the groups defined by the random effect components. Specifically, for $G = 1, \dots, 10$ we compute the Normalized Entropy Criterion (NEC)

$$NEC_G = \frac{EN_G}{\ell(\mathbf{\Phi};G) - \ell(\mathbf{\Phi};1)}$$

Celeux and Soromenho (1996), where $\ell(\mathbf{\Phi};G)$ denotes the log-likelihood of the model with G components and, letting $\tilde{\pi}_{ik} = P(u_i = u_k \mid y_{i1}, \dots, y_{in_i}, \boldsymbol{x}_{i1}, \dots, \boldsymbol{x}_{in_i})$,

$$EN_G = -\frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{G} \tilde{\pi}_{ik} \log \tilde{\pi}_{ik}$$

is an average entropy measure. As with AIC and BIC, lower values of NEC should be preferred. Notice that NEC_1 is not defined, being conventionally set to 1.

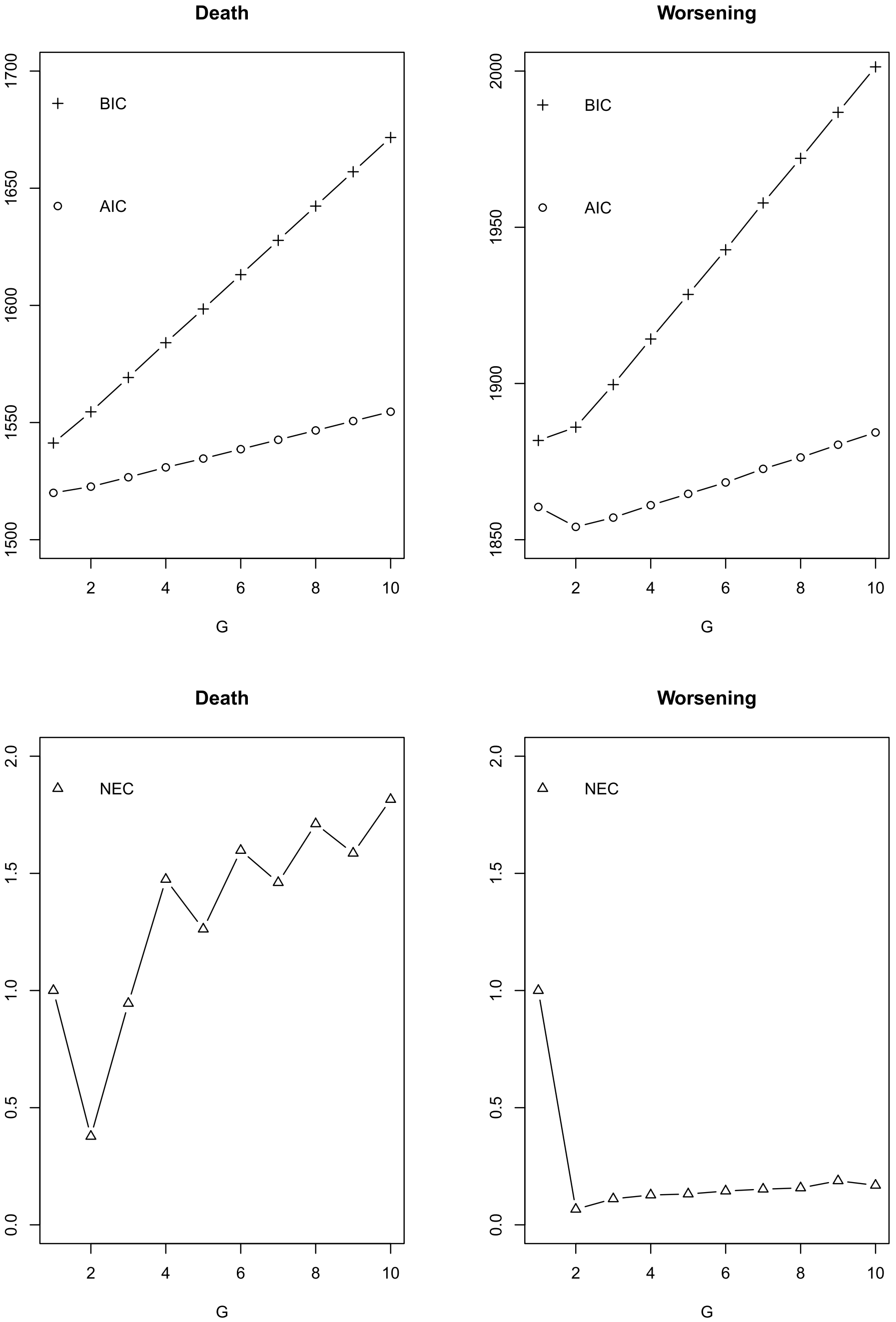


Fig. 2 AIC, BIC and NEC of FMLMs for the outcome death (left panel) and worsening (right panel)

Table 4 FMLM: parameter estimates for the outcomes Death and Worsening. Standard errors of estimates are in parentheses (significance codes: '.' 0.1, '*' 0.05, '**' 0.01, '***' 0.001)

	Death		Worsening	
Gender (male)	0.235		0.199	
	(0.155)		(0.136)	
Age	0.042	***	0.035	***
	(0.007)		(0.006)	
IRW	1.624	***	0.581	**
	(0.238)		(0.217)	
Mass 1	−6.397	***	−4.462	***
	(0.708)		(0.596)	
Mass 2	−5.742	***	−3.788	***
	(0.720)		(0.594)	
	Mixture Proportions			
Location 1	0.879		0.747	
Location 2	0.121		0.253	

Figure 2 displays AIC, BIC and NEC as a function of G for the outcome Death (left panel) and Worsening (right panel). The three criteria agree on selecting $G = 2$ components for the outcome variable Worsening. For the outcome Death, on the other hand, AIC and BIC do not find evidence of a NH effect ($G = 1$), while NEC would lead to select $G = 2$ components. To further investigate the structure of the data and the shape of the NH distribution, we look at parameter estimates and NH effect for $G = 2$.

Table 4 reports the parameter estimates of the models for the two outcome variables when $G = 2$. It can be noted that the parameter estimates for the covariates with fixed effects are very similar to those reported in Table 3. Each of the two mixture components is identified by a location of the random intercept (Location 1 and 2) in the Table: the larger value implies that the probability of Death or Worsening is higher, conditional on the values of the explicative variables. This means that, given two residents with the same values of the explicative variables, the largest intercept yields a greater chance of Death or Worsening than the lowest one. So, the mixture component for which $k = 2$ identifies the NHs with a higher chance of Death or Worsening given the Age, Gender and IRW of the resident.

Now, for each NH, the posterior probabilities of belonging to the first and to the second component of the mixture can be computed with Eq. (6). Using these probabilities, the posterior expected value of the random intercept can be computed as well using Eq. (7). Then, we can compare the distributions of these quantities to those obtained with the LMM in the previous section. Figure 3 reports the prior (panel a) and the posterior (panel b) distribution of u_i's under the FMLM and the posterior distribution of u_i's (panel c) for the LMM, after centering the random effects on 0 for comparison. The plots are very similar for the two outcomes. We can note the presence of a smaller mass that deviates from the larger one, thus highlighting the presence of a small group of NHs

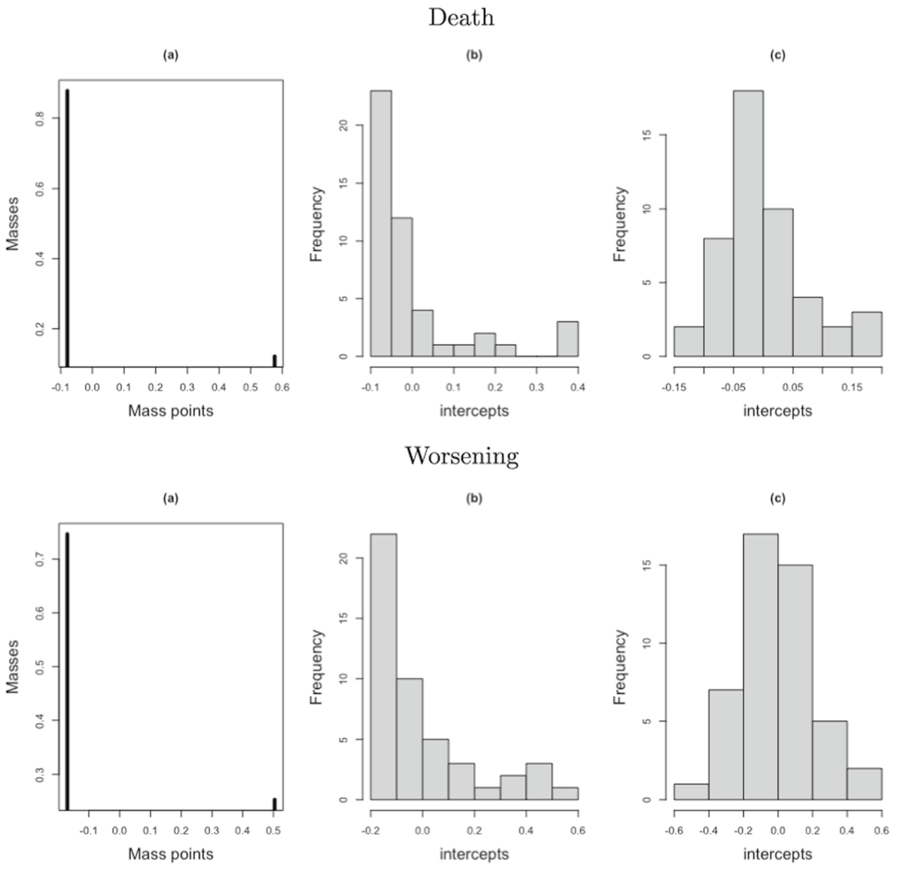


Fig. 3 FMLM: estimated prior (**a**) and posterior (**b**) distribution for u_i's. LMM: posterior distribution for u_i's (**c**)

with a higher probability of Death and of Worsening. Looking at the posterior probabilities of belonging to the first and to the second component, the latter is the modal one for NHs 6, 12 and 34 with respect to the outcome Death and for NHs 1, 6, 8, 9, 10, 11, 12, 23, and 31 with respect to the outcome Worsening. This can be a useful tool for assessing NH performance and, in particular, to detect particularly alarming NHs.

In panels (b) of Fig. 3 we can see the distribution of the posterior intercepts obtained from the mixture and in panels (c) the histogram obtained by assuming Normality in the LMM. This comparison is particularly useful as a diagnostic tool for normality of random effects. It is clear from plots (b) that the distribution of the random effects is heavily skew in both cases and that the assumption of Normality for them is not suitable in this context. Note that the asymmetry of the distribution of the NH effects remains also as the number of mixture components increases, as it is evident in Figs. 6 and 7 in the Appendix where the a-priori distribution for all

Table 5 M-quantile model: fixed effects for binary response Death and Worsening. Standard errors of estimates are in parentheses (significance codes: '.' 0.1, '*' 0.05, '**' 0.01, '***' 0.001)

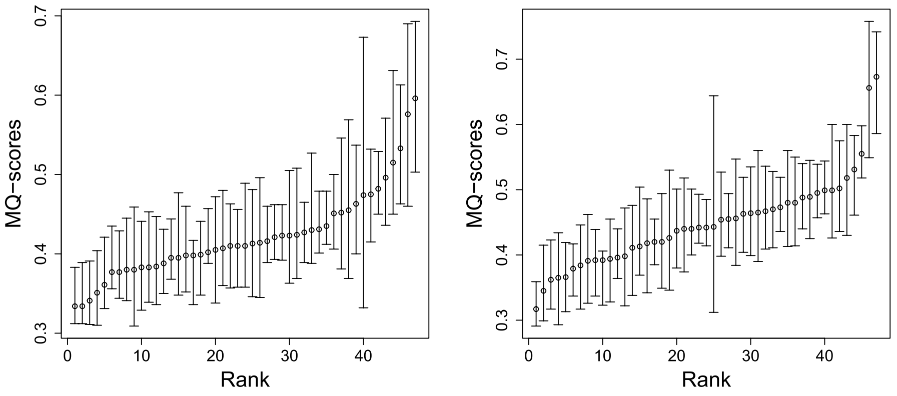
	Death	Worsening
Intercept	-6.163^{***}	-4.199^{***}
	(0.720)	(0.591)
Gender (male)	0.212	0.204
	(0.156)	(0.134)
Age	0.041^{***}	0.034^{***}
	(0.008)	(0.006)
IRW	1.532^{***}	0.566^{***}
	(0.239)	(0.215)

values of G and for both responses are reported. For these reasons in the next section we propose a more robust approach to evaluate the performance of the NHs.

4.3 Results from M-quantile regression models

M-Quantile regression models for binary data are fitted to the two outcome variables—Death and Worsening—using as covariates the variables Gender, Age and IRW. Therefore, NH membership is not included in the model either as a fixed or as a random effect. In fact, as discussed above, here we take a different perspective and evaluate the presence and magnitude of a NH effect directly from the data without assuming a distribution and looking at where units lay at the different levels (M-quantiles) of the conditional distribution of the response. We fit the models using ad-hoc functions developed by the Authors in R and available from the Authors upon request. In order to estimate M-quantile coefficients q_{ij}'s, 197 equally spaced values of q between 0.01 and 0.99 are used. Table 5 reports the estimates of the fixed effects of the covariates for $q = 0.5$, while Figs. 8 and 9 in the Appendix show the estimated regression coefficients for the covariates for each value of q with the corresponding 95% confidence bounds for the outcome Death and for Worsening, respectively. Standard errors and confidence bounds are obtained by means of 5000 bootstrap samples drawn using a nonparametric block-bootstrap approach that resamples with replacement within each NH. The value of the tuning constant c is set to 1.345 as it is customary in many applications (see e.g. Chambers et al. 2016). Again, very similar results are obtained for the effect of the covariates. Recall that for $q = 0.5$ and $c \to \infty$ M-quantile regression is equivalent to logistic regression. Note that for this data there is no evidence of M-quantile crossing, and this does not provide evidence of model misspecification.

To measure the NH effects on the probabilities of Death or of Worsening, the following procedure has been implemented. For every resident, the M-quantile coefficient is computed by means of Eq. (14) using the fine grid of q values. Then, by averaging the M-quantile coefficients of the residents in the same NH, we obtain an average score for each NH and we call it MQ-score. The larger the value of this MQ-score for a NH, the higher the probability of Death or of



(a) *Death*

(b) *Worsening*

Fig. 4 MQ-score of NHs for Death (**a**) and Worsening (**b**) and corresponding bootstrap based pairwise overlap intervals. NHs are arranged by the rank of the MQ-score

Table 6 NH MQ-scores and pairwise overlap interval limits for the binary responses Death and Worsening: first and last six positions in the ranking

Death				*Worsening*			
NHs	Est	low	upper	NHs	Est	low	upper
4	0.334	0.312	0.383	4	0.317	0.291	0.359
16	0.334	0.312	0.389	16	0.345	0.299	0.415
40	0.341	0.311	0.391	41	0.362	0.317	0.423
36	0.351	0.310	0.404	47	0.365	0.293	0.434
18	0.361	0.331	0.421	35	0.366	0.313	0.419
32	0.377	0.356	0.435	33	0.379	0.337	0.417
...	...	...	...				
6	0.482	0.450	0.529	34	0.502	0.436	0.575
17	0.496	0.436	0.571	23	0.518	0.430	0.600
45	0.515	0.450	0.631	9	0.531	0.461	0.583
34	0.533	0.463	0.613	6	0.555	0.518	0.598
10	0.576	0.460	0.690	10	0.656	0.549	0.758
12	0.596	0.503	0.693	12	0.673	0.586	0.742

Worsening for its residents, conditional on their values of the covariates. So, the MQ-score allows to compare the NHs with respect to the ability of avoiding the decease of the resident or the worsening of his/her health conditions taking into account the NH case-mix.

Figure 4 shows the MQ-scores of the NHs ordered by size with the corresponding block-bootstrap overlap intervals for pairwise comparisons (Goldstein and Healy 1995). When the intervals of two NHs do not overlap, their MQ-score difference is significant at approximately a 5% level of the first type error. The MQ-score ranges from a minimum of 0.334 to a maximum of

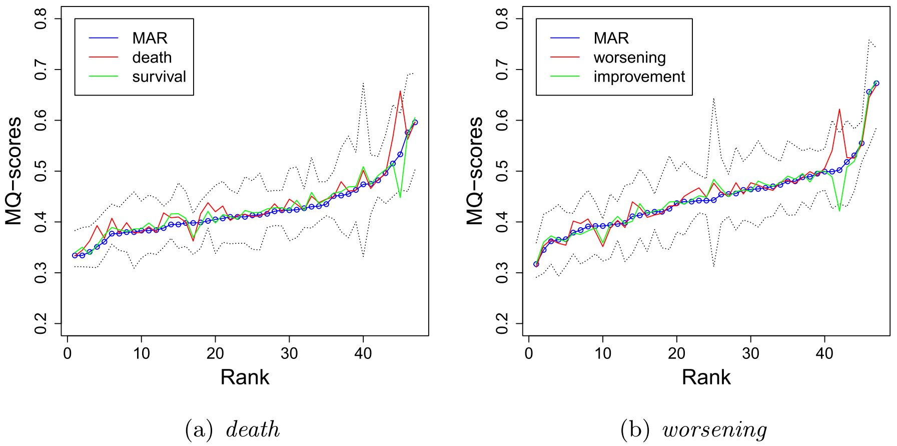
0.596 for the response Death, and from a minimum of 0.317 to a maximum of 0.673 for the response Worsening. Table 6 reports details of the first and last six positions in the ranking with respect to the MQ-score of the NHs for the two binary responses. All pairwise comparisons between the last six and the first six positions in the rankings are significant, apart from the 23 to 47 pair in the ranking for Worsening.

Note that using the FMLM approach, the NHs classified in the second group (larger probability of Death or of Worsening) also have a very large MQ-score. In fact, for the Death outcome we find NHs 6, 12 and 34 in the last six positions of Table 6; while for the outcome Worsening NHs 6, 9, 10, 12 and 23 are the worst with respect to the posterior expected value of the random effect in the FMLM and with respect to the MQ-score as well.

Finally, the average M-quantile coefficient of all residents is 0.427 for the response Death and 0.455 for the response Worsening. Taking these values as fixed and considering the 95% bootstrap confidence intervals for the NH MQ-scores, NHs 4, 16 and 40 are significantly below the average M-quantile coefficient, while NHs 6, 12 and 34 are significantly above it for the response Death; for the response Worsening, NHs 4, 16, 33, 35, 39, and 41 are significantly below the average overall M-quantile coefficient, while NHs 6, 10, and 12 are significantly above it. In this way it is possibile to single out instances of best practices or cases of unsatisfactory performances.

4.4 Sensitivity analysis for residents' discharge

Until now, the response values that are missing because of discharge of residents have been taken as MAR, as no information is available in the data on the reasons of



(a) *death* (b) *worsening*

Fig. 5 MQ-scores of NHs obtained under the MAR assumption (blu circles), under the assumption that discharges are all deceased or worsened (red lines), and under the assumption that discharges are all survived or not worsened (green lines). NHs are arranged by the rank of the MQ-score under the MAR assumption

the discharge. In this section we perform a sensitivity analysis to evaluate the impact of discharges on the estimates and the robustness of the results quoted in the previous sections. To this end, for the response Death we consider two scenarios: in the first one, a discharge is assumed to correspond to a death; for example, the NH resident is discharged because he/she needs hospitalization or a level of care that the NH is not able to provide; in the second setting, a discharge is assumed to correspond to a survival, as when there is an improvement that no longer requires staying in the NH. Similarly for the response Worsening, in the first setting a discharge is considered as a worsening, while in the second setting as a non-worsening.

Clearly, without further information, the two extreme scenarios discussed above are not necessarily a good approximation of the reality for the two outcomes, as the MAR assumption might not be either. Nevertheless, these scenarios provide two patterns that can be compared to the findings obtained under MAR. In this way, the overall validity of the latter can be assessed to some extent. With this regard, Fig. 5 shows the behavior of the NH MQ-scores ordered under the MAR assumption (blue circle line), with the limit of the pairwise comparison overlap intervals (dotted line). In addition, the red line shows the corresponding values under the first setting (discharge as death on the left or worsened on the right), while the green line shows the corresponding values under the second setting (discharge as survival on the left or improvement on the right). Now, the blue, red and green lines follow similar patterns and are all within the confidence bounds for all NHs but NH 34. In fact, NH 34 has the highest rate of discharged residents, and uncertainty on its performance may require attention by the regional health management.

5 Discussion and conclusions

In this paper we focus on the evaluation of the performance of institutions delivering services. The aim is to develop tools that allow monitoring the institutions, compare them from an efficacy point of view, and inform policy makers. In this regard we use statistical models for evaluating the performance of institutions with respect to outcome variables that are dichotomous, taking into account user case-mix adjustments to allow fair comparisons. Using data regarding users of the services delivered by the institutions, it is possibile to estimate the contribution to the outcomes of the institutions separated by that of the case-mix. Measuring this contribution would allow the ranking of the institutions from the most effective to the least one and single out best practices or cases of bad services. In particular, we review the classical approach that uses logistic mixed models and investigate robust alternatives such as finite mixtures of logistic models and M-Quantile regression for binary data.

The application refers to Nursing Homes (NHs) for elderly that need long term health care. For these institutions, ability to keep residents healthy is of interest and in this respect the outcome variables considered in this work are binary indicators of decease and of worsening over a given period of time. To adjust for the case-mix of residents of each NH, the use of the Resource Utilization Group (RUG) system as proxy of the health conditions is proposed. In particular, the RUG system classifies residents in homogeneous groups according to the kind of impairment of the elderly and the treatments received. To each RUG group a weight is attached as an indicator of the amount of care burden it requires.

The results obtained from fitting the statistical models discussed in the paper to the data from the NHs of the Italian region Umbria show that the most relevant predictive variable of the outcomes within the RUG framework is the weight of the RUG group to which each resident is assigned. Beside that, significant effects are shown for age and gender of residents. Conditional on the latter, we were interested in finding a NH effect on the outcomes which can be interpreted as its ability of avoiding the death or the worsening of the resident.

Using Finite Mixtures of Logistic Models, it has been possible to check that the assumption of Gaussian random effects used in logistic mixed models is not appropriate for the data at hand, so that a more robust approach based on M-Quantile regression for binary data has been proposed and effectively applied to obtain a score for each NH (MQ-scores). Similarly to random effects from logistic mixed models, the MQ-scores allow to rank NHs from the most effective to the least effective in delaying the deterioration of the health conditions of the elderly. However, no parametric assumption for the distribution of these effects has been made and, in addition, possible interactions of the NH group effect with the covariates is automatically accounted for. A nonparametric block-bootstrap procedure has been used to measure the standard error of the MQ-scores and build confidence and overlap intervals to make comparisons among NHs. Since M-quantile regression for binary data is a direct expectile-like extension of logistic regression, we believe that MQ-scores provide a valuable alternative to logistic mixed models for performance evaluation that is data driven and robust.

Fair comparisons between NHs with respect to the outcomes of decease or worsening, especially when they are statistically significant, can be used for the choice of the facility, or for identifying best practices or cases of mismanagement or unacceptable level of the care. Results obtained using the MQ-score are in line with the clustering obtained using finite mixtures of logistic models. These two robust approaches have been proved to be useful alternatives to classical mixed effects models with Gaussian random effects to evaluate the performance of institutions. The extension of both approaches to incorporate longitudinal data is a topic for further research. Similarly, the paper by Alfò et al. (2017) considers finite mixtures of linear M-quantile regression models and it would be interesting to extend their approach to M-quantile regression for binary data. Also, an alternative method for computing the M-quantile coefficient for binary data has been recently developed by Dawber et al. (2022) using expectile regression. Application of this alternative approach for evaluation purposes is also a topic for further research. Finally, the paper has focused on modeling the conditional expectation of the binary response and its extensions in an expectile-like fashion. In this respect, it is of interest to investigate possible extensions of new approaches in the quantile regression literature, such as that proposed by Geraci and Farcomeni (2022) for count data, for evaluation problems based on binary outcomes as the one faced here.

Supplementary figures

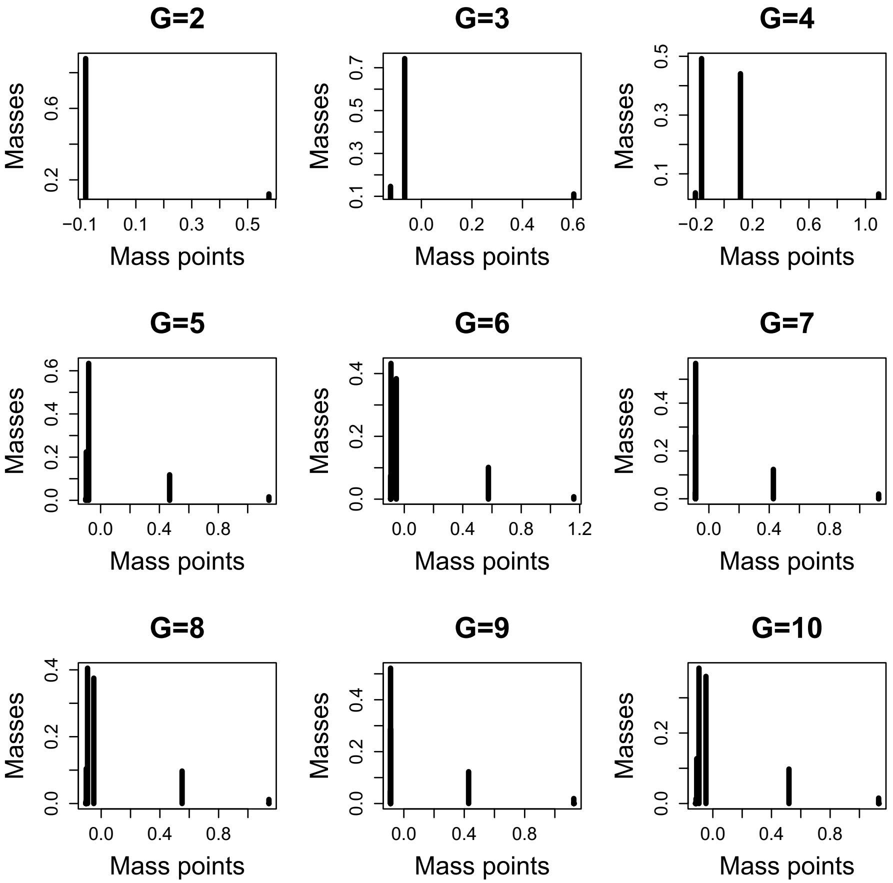


Fig. 6 Prior distribution of the random effects for a number of mass points (*G*) ranging from 2 to 10. Death outcome

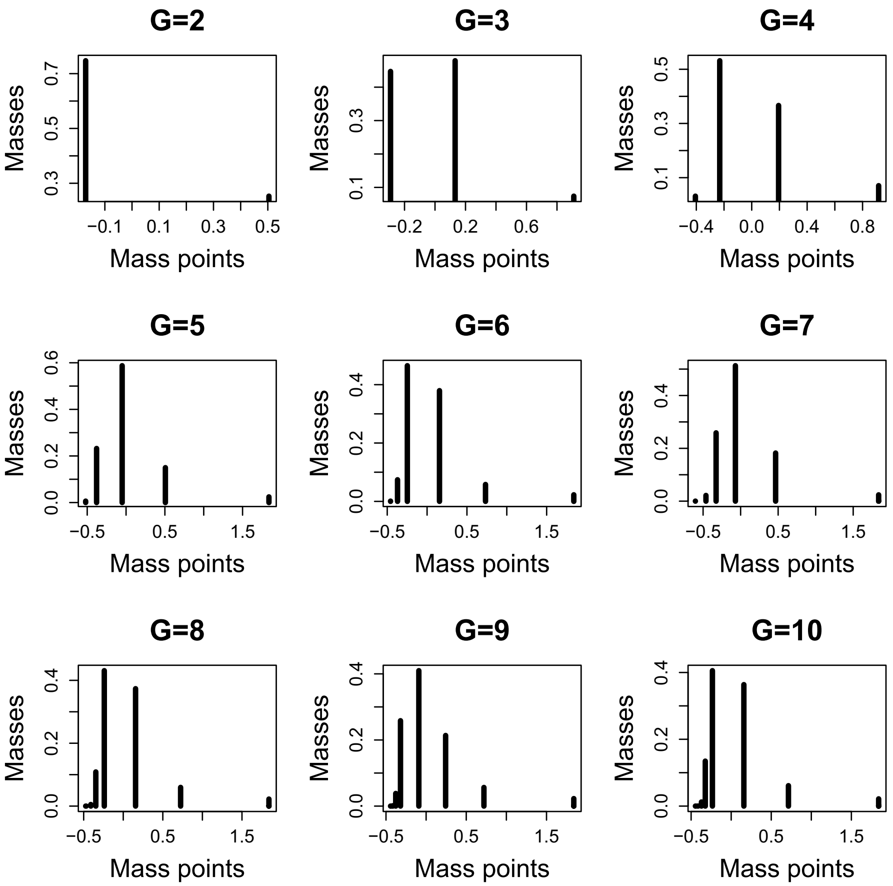


Fig. 7 Prior distribution of the random effects for a number of mass points (G) ranging from 2 to 10. Worsening outcome

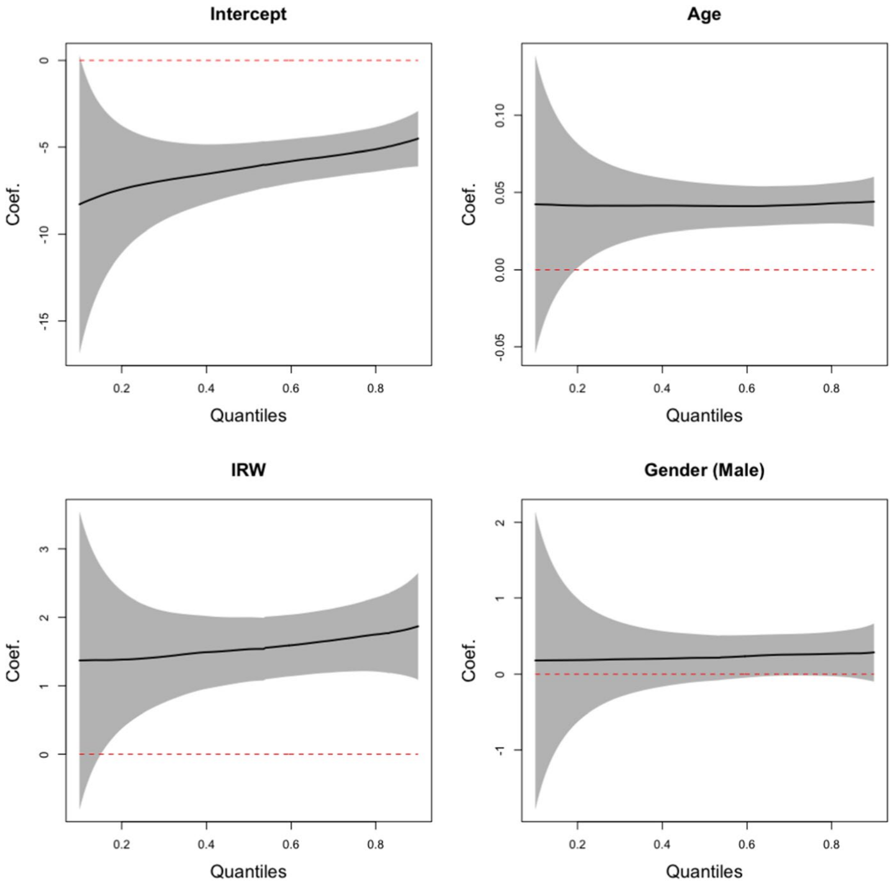


Fig. 8 M-Quantile regression coefficient estimates with respect to q with bootstrap based confidence bounds. Death outcome

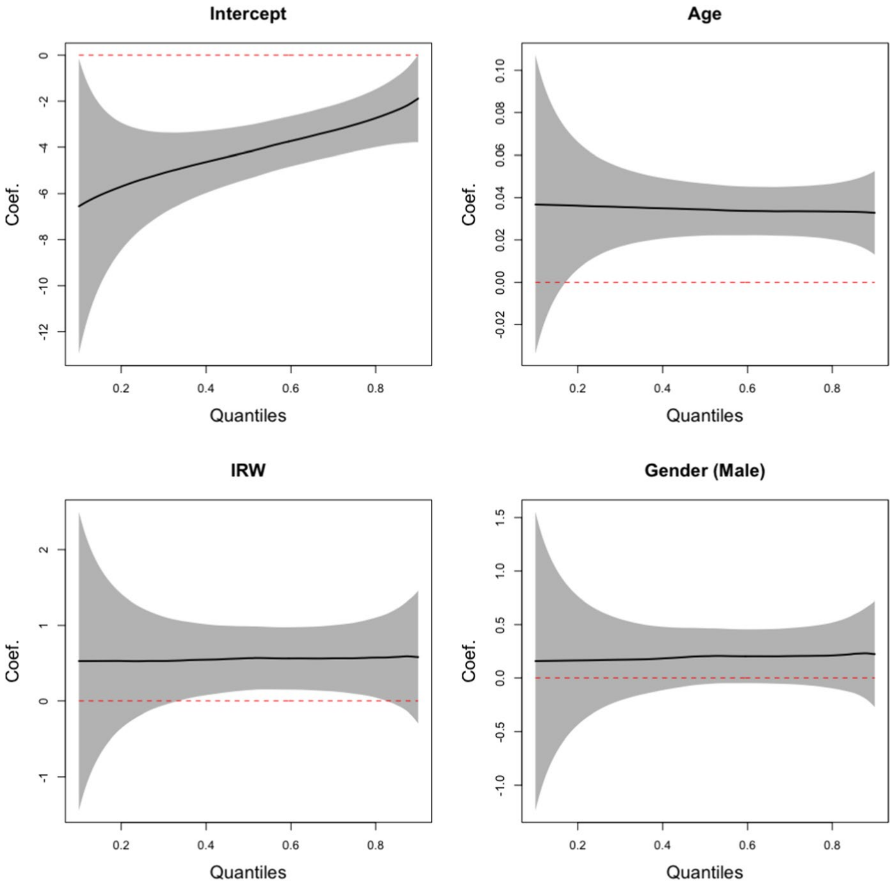


Fig. 9 M-Quantile regression coefficient estimates with respect to q with bootstrap based confidence bounds. Worsening outcome

Acknowledgements The work of De Novellis, Doretti, Montanari and Ranalli is supported by Fondazione Cassa di Risparmio di Perugia (grant 2018.0420). The work of Salvati has been developed under the support of the project PRIN-Quantification in the Context of Dataset Shift - QuaDaSh (Grant P2022TB5JF, Italy). The work of De Novellis and Doretti was conducted while they were PhD student and Assistant Professor, respectively, at the University of Perugia. The Authors are grateful to Regione Umbria (Italy) for sharing the data that inspired this work.

Funding Open access funding provided by Università degli Studi di Perugia within the CRUI-CARE Agreement. This research is supported by Fondazione Cassa di Risparmio di Perugia (grant 2018.0420) and by project PRIN-Quantification in the Context of Dataset Shift - QuaDaSh (Grant P2022TB5JF, Italy).

Declarations

Conflict of interest All authors declare that they have no Conflict of interest.

References

Aitkin M (1996) A general maximum likelihood analysis of overdispersion in generalized linear models. Statistics and Computing 6:251–262

Aitkin M (1999) A general maximum likelihood analysis of variance components in generalized linear models. Biometrics 55:117–128

Alfò M, Salvati N, Ranalli MG (2017) Finite mixtures of quantile and M-quantile regression models. Statistics and Computing 27:547–570

Aristodemou K, He J, Yu K (2019) Binary quantile regression and variable selection: A new approach. Econometric Reviews 38(6):679–694

Bartolucci F, Lupparelli M, Montanari GE (2009) Latent Markov model for longitudinal binary data: An application to the performance evaluation of nursing homes. The Annals of Applied Statistics 3(2):611–636

Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. Journal of Statistical Software 67(1):1–48

Benoit DF, Van den Poel D (2012) Binary quantile regression: a bayesian approach based on the asymmetric laplace distribution. Journal of Applied Econometrics 27(7):1174–1188

Berlowitz DR, Ash AS, Brandeis GH, Brand HK, Halpern JL, Moskowitz MA, Gwaltney JM Jr (1996) Rating long-term care facilities on pressure ulcer development: importance of case-mix adjustment. Annals of Internal Medicine 124(6):557–563

Berta P, Ingrassia S, Punzo A, Vittadini G (2016) Multilevel cluster-weighted models for the evaluation of hospitals. Metron 74(3):275–292

Berta P, Vinciotti V (2019) Multilevel logistic cluster-weighted model for outcome evaluation in health care. Statistical Analysis and Data Mining: The ASA Data Science Journal 12(5):434–443

Breckling J, Chambers R (1988) M-quantiles. Biometrika 75:761–771

Breslow NE, Clayton DG (1993) Approximate inference in generalized linear mixed models. Journal of the American Statistical Association 88(421):9–25

Brizioli E, Bernabei R, Grechi F, Masera F, Landi F, Bandinelli S, Cavazzini C, Gangemi S, Ferrucci L (2003) Nursing home case-mix instruments: Validation of the RUG-III system in Italy. Aging Clinical and Experimental Research 15(3):243–253

Broussard DM, Reiter KL (2020) Estimated reduction in CAH profitability from loss of cost-based reimbursement for swing beds. Technical report, North Carolina Rural Health Research Program

Cantoni E, Ronchetti E (2001) Robust inference for generalized linear models. Journal of the American Statistical Association 96:1022–1030

Carpenter I, Hirdes J P (2013) Using interRAI assessment systems to measure and maintain quality of long-term care. In A Good Life in Old Age? Monitoring and Improving Quality in Long-Term Care, chapter 3, pages 93–139. OECD Health Policy Studies

Celeux G, Soromenho G (1996) An entropy criterion for assessing the number of clusters in a mixture model. Journal of Classification 13:195–212

Chambers R, Salvati N, Tzavidis N (2016) Semiparametric small area estimation for binary outcomes with application to unemployment estimation for local authorities in the UK. Journal of the Royal Statistical Society. Series A (Statistics in Society), pages 453–479

Dawber J, Salvati N, Fabrizi E, Tzavidis N (2022) Expectile regression for multi-category outcomes with application to small area estimation of labour force participation. Journal of the Royal Statistical Society. Series A (Statistics in Society), page https://doi.org/10.1111/rssa.12953.

Dempster A, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B 39:1–38

Dreassi E, Ranalli MG, Salvati N (2014) Semiparametric M-quantile regression for count data. Statistical Methods in Medical Research 23(6):591–610

Einbeck J, Darnell R, Hinde J (2018) npmlreg: Nonparametric Maximum Likelihood Estimation for Random Effect Models. R package version 0.46-5

Fiaschi D, Giuliani E, Nieri F, Salvati N (2020) How bad is your company? Measuring corporate wrongdoing beyond the magic of esg metrics. Business Horizons 63(3):287–299

Fries BE, Schneider DP, Foley WJ, Gavazzi M, Burke R, Cornelius E (1994) Refining a case-mix measure for nursing homes: Resource Utilization Groups (RUG-III). Medical Care 32(7):668–685

Gasperoni F, Ieva F, Paganoni AM, Jackson CH, Sharples L (2020) Non-parametric frailty cox models for hierarchical time-to-event data. Biostatistics 21(3):531–544

Geraci M, Farcomeni A (2022) Mid-quantile regression for discrete responses. Statistical Methods in Medical Research 31(5):821–838

Goldstein H (2011) Multilevel statistical models. John Wiley & Sons

Goldstein H, Healy MJR (1995) The graphical presentation of a collection of means. Journal of the Royal Statistical Society. Series A 158(1):175–177

Grieco N, Ieva F, Paganoni AM (2012) Performance assessment using mixed effects models: a case study on coronary patient care. IMA Journal of Management Mathematics 23(2):117–131

Hirdes JP, Ljunggren G, Morris JN, Frijters DH, Finne Soveri H, Gray L, Björkgren M, Gilgen R (2008) Reliability of the interRAI suite of assessment instruments: A 12-country study of an integrated health information system. BMC Health Services Research 8:277

Huber P (1973) Robust regression: Asymptotics, conjectures and Monte Carlo. The Annals of Statistics 1:799–821

Ikegami N, Fries BE, Takagi Y, Ikeda S, Ibe T (1994) Applying RUG-III in Japanese long-term care facilities. The Gerontologist 34(5):628–639

Jiang J (1998) Consistent estimators in generalized linear mixed models. Journal of the American Statistical Association 93:720–729

Kim H, Jung Y-I, Sung M, Lee J-Y, Yoon J-Y, Yoon J-L (2015) Reliability of the interRAI Long Term Care Facilities (LTCF) and interRAI Home Care (HC). Geriatrics & Gerontology International 15:220–228

Koenker R, Bassett G (1978) Regression quantiles. Econometrica 46:33–50

Kokic P, Chambers R, Breckling J, Beare S (1997) A measure of production performance. Journal of Business & Economic Statistics 15(4):445–451

Kordas G (2006) Smoothed binary regression quantiles. Journal of Applied Econometrics 21(3):387–407

Laird N (1978) Nonparametric maximum likelihood estimation of a mixing distribution. Journal of the American Statistical Association 73:805–811

Lindsay BG (1983) The geometry of mixture likelihoods: a general theory. The Annals of Statistics 11:86–94

Lindsay BG (1983) The geometry of mixture likelihoods, Part II: the exponential family. The Annals of Statistics 11:783–792

Little R J, Rubin D B (2002) Statistical analysis with missing data: Wiley series in probability and statistics. Wiley (New York, NY)

Masci C, Ieva F, Paganoni AM (2022) Semiparametric multinomial mixed-effects models: A university students profiling tool. The Annals of Applied Statistics 16(3):1608–1632

Masci C, Paganoni AM, Ieva F (2019) Semiparametric mixed effects models for unsupervised classification of italian schools. Journal of the Royal Statistical Society Series A: Statistics in Society 182(4):1313–1342

McCulloch CE (1997) Maximum likelihood algorithms for generalized linear mixed models. Journal of the American Statistical Association 92:162–170

Montanari GE, Doretti M (2019) Ranking nursing homes' performances through a latent markov model with fixed and random effects. Social Indicators Research 146(1):307–326

Montanari GE, Doretti M, Bartolucci F (2018) A multilevel latent Markov model for the evaluation of nursing homes' performance. Biometrical Journal 60(5):962–978

Newey W, Powell J (1987) Asymmetric least squares estimation and testing. Econometrica 55:819–847

Pinheiro J, Bates D (1995) Approximations to the log-likelihood function in the nonlinear mixed-effects model. Journal of Computational and Graphical Statistics 4:12–35

Pinheiro J, Bates D (2000) Mixed-effects models in S and S-PLUS. Springer science & business media

Punelli D, Williams S (2013) Nursing facility reimbursement and regulation. Technical report, Research Department, Minnesota House of Representatives

Ragni A, Masci C, Ieva F, Paganoni A M (2023) Clustering hierarchies via a semi-parametric generalized linear mixed model: a statistical significance-based approach. arXiv preprint arXiv:2302.12103

Saei A, Chambers R (2003) Small area estimation under linear and generalized linear mixed models with time and area effects. In S3RI Methodology Working Papers, pages 1–35. Southampton Statistical Sciences Research Institute, Southampton

Schneider D P, Fries B E, Foley W J, Desmond M, Gormley W J (1988) Case mix for nursing home payment: resource utilization groups, version II. Health Care Financing Review, pages 39–52

Simar L (1976) Maximum likelihood estimation of a compound Poisson process. The Annals of Statistics 4:1200–1209

Tzavidis N, Ranalli MG, Salvati N, Dreassi E, Chambers R (2015) Robust small area prediction for counts. Statistical Methods in Medical Research 24(3):373–395

Tzavidis N, Salvati N, Schmid T, Flouri E, Midouhas E (2015) Longitudinal Analysis of the Strengths and Difficulties Questionnaire Scores of the Millennium Cohort Study Children in England Using M-Quantile Random-Effects Regression. Journal of the Royal Statistical Society Series A: Statistics in Society 179(2):427–452

Wedel M, DeSarbo WS, Bult JR, Ramaswamy V (1993) A latent class Poisson regression model for heterogeneous count data. Journal of Applied Econometrics 8:397–411

Wray NP, Hollingsworth JC, Petersen NJ, Ashton CM (1997) Case-mix adjustment using administrative databases: A paradigm to guide future research. Medical Care Research and Review 54(3):326–356

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.